

Capstone Project

Hotel Booking Analysis

- Kanishque Tyagi

Problem Statement:

- **Give data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of required parking spaces, among other things.**
- **To Explore and analyze the data to discover important factors that govern the bookings.**

Contents:

S. No	Topic	Page
1	Columns Description	4
2	Data Description	5
3	Data Cleaning	6
4	Distribution of Bookings	7-9
5	Stay Duration	10-11
6	Geographical Analysis	12-13
7	ADR Analysis	15
8	Revenue Analysis	16
9	Customer Analysis	17-18
10	Market segment & Distribution Channel	19
11	Cancellation Analysis	20-25

Columns Names and Their Meanings:

- **hotel**: hotel type(H1 = Resort Hotel or H2 = City Hotel)
- **is_canceled**: Value indicating if the booking was canceled (1) or not (0)
- **lead_time**: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- **arrival_date_year**: Year of arrival date
- **arrival_date_month**: Month of arrival date
- **stays_in_weekend_nights and stays_in_week_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **traveler's** : Number of adults + Number of children + Number of babies
- **Canceled** – booking was canceled by the customer
- **adr**: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights. (measures the average rental revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using the ADR. Multiplying the ADR by the occupancy rate equals the revenue per available room.)
- **country**: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- **market_segment**: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- **is_repeated_guest**: Value indicating if the booking name was from a repeated guest (1) or not (0)
- **room_type**: Code of room type. Code is presented instead of designation for anonymity reasons.
- **deposit_type**: Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
 - No Deposit – no deposit was made
 - Non Refund – a deposit was made in the value of the total stay cost
 - Refundable – a deposit was made with a value under the total cost of stay.
- **days_in_waiting_list**: Number of days the booking was in the waiting list before it was confirmed to the customer.
- **customer_type**: Type of booking, assuming one of four categories:
 - Contract - when the booking has an allotment or other type of contract associated to it
 - Group – when the booking is associated to a group
 - Transient – when the booking is not part of a group or contract, and is not associated to other transient booking

Data Description:

```
# dataset summary statistics - Integer and numeric variables.  
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

Numerical Features

```
# dataset summary statistics - Categorical variables.  
df[discriptive_cols].describe().transpose()
```

	count	unique	top	freq
hotel	119390	2	City Hotel	79330
arrival_date_month	119390	12	August	13877
meal	119390	5	BB	92310
country	118902	177	PRT	48590
market_segment	119390	8	Online TA	56477
distribution_channel	119390	5	TATO	97870
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
customer_type	119390	4	Transient	89613
reservation_status	119390	3	Check-Out	75166
reservation_status_date	119390	926	2015-10-21	1461

Categorical Features

Data Cleaning:

There are four columns having Null values:

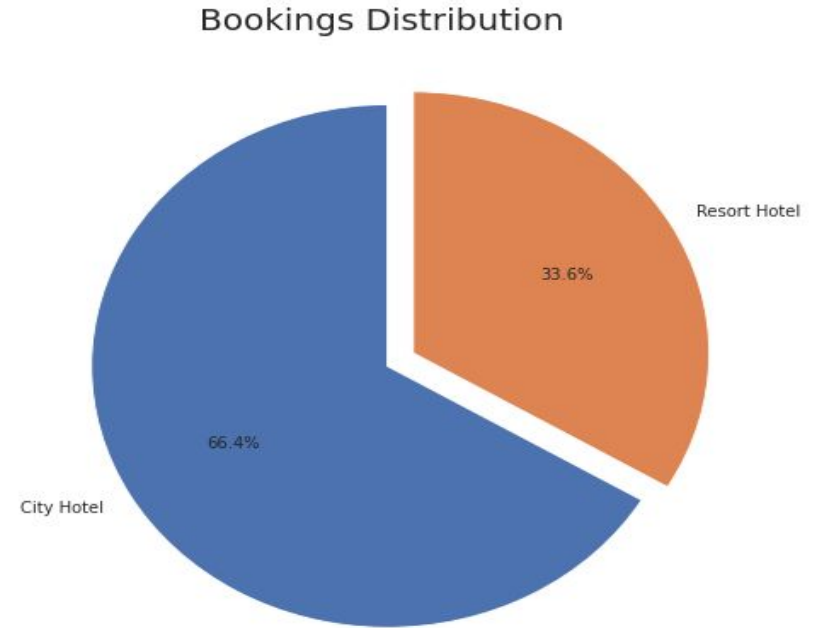
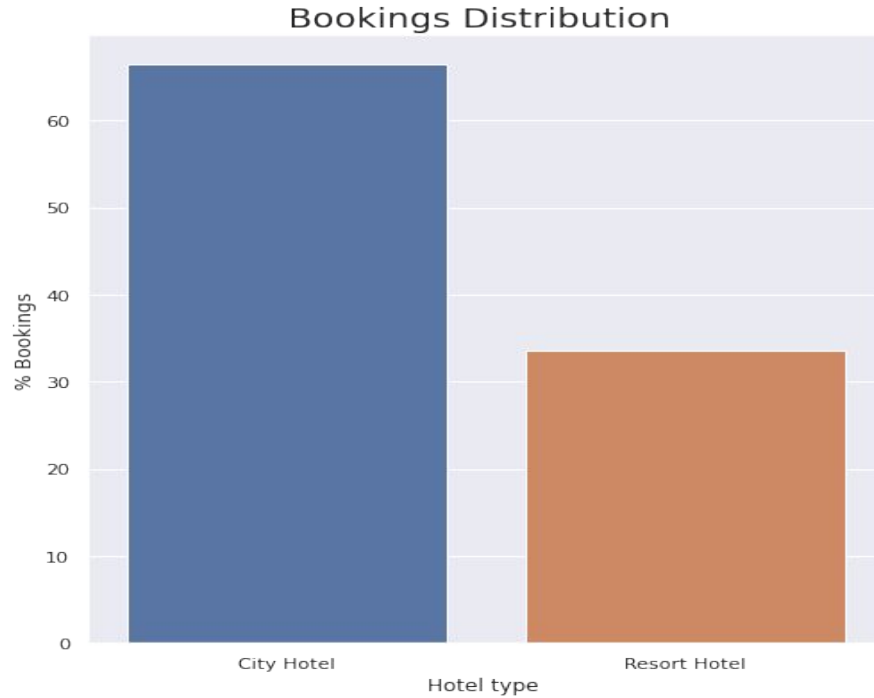
1. Company - 112593
2. Agent - 16340
3. Country - 488
4. Children - 4

Dealing with null values:

- Filled Null values in children and agent columns with median values of respective columns.
- For country column, filled mode in place of Null values.
- Dropped the Company column since it has more than 80 % Null Values.

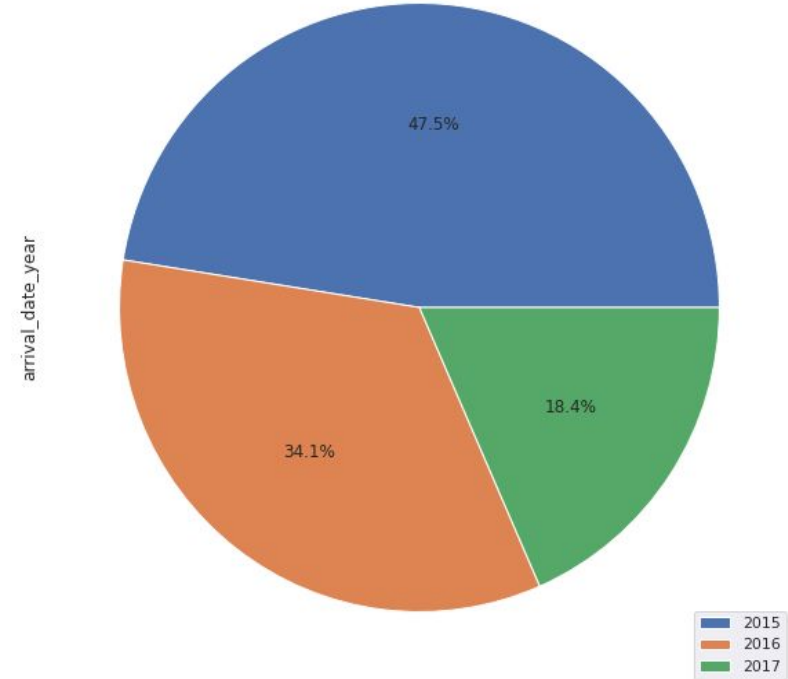
After these Steps Data became ready to analyze.

Distribution of Booking by Hotel Type



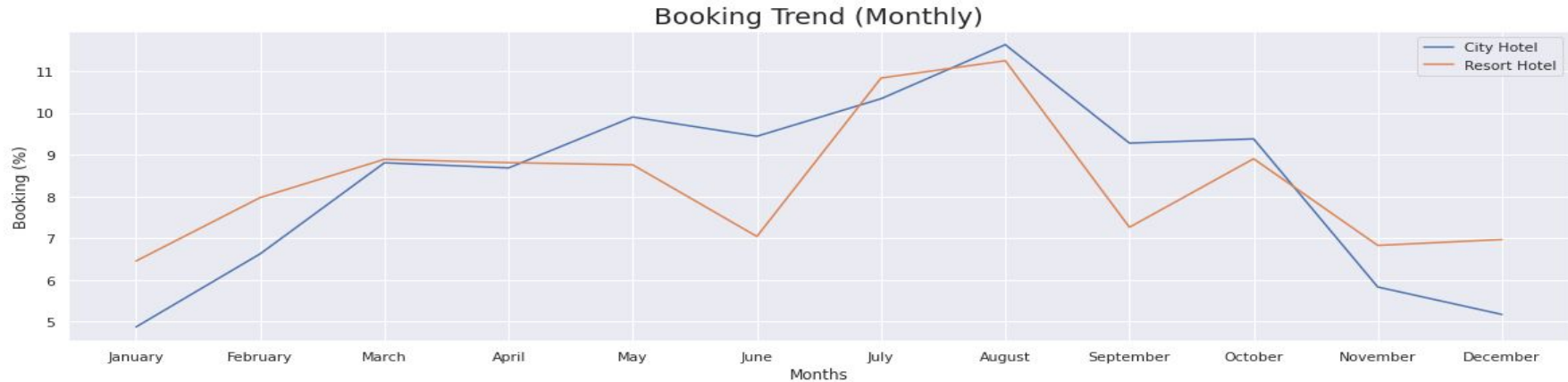
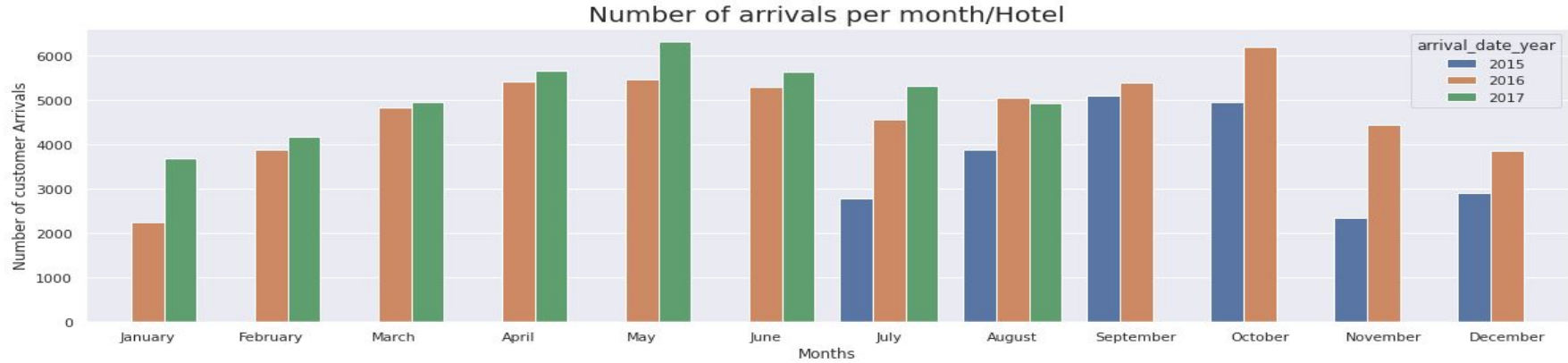
City Hotel Has Higher Booking Percentage Than Resort Hotel

Distribution of booking by Year-wise:



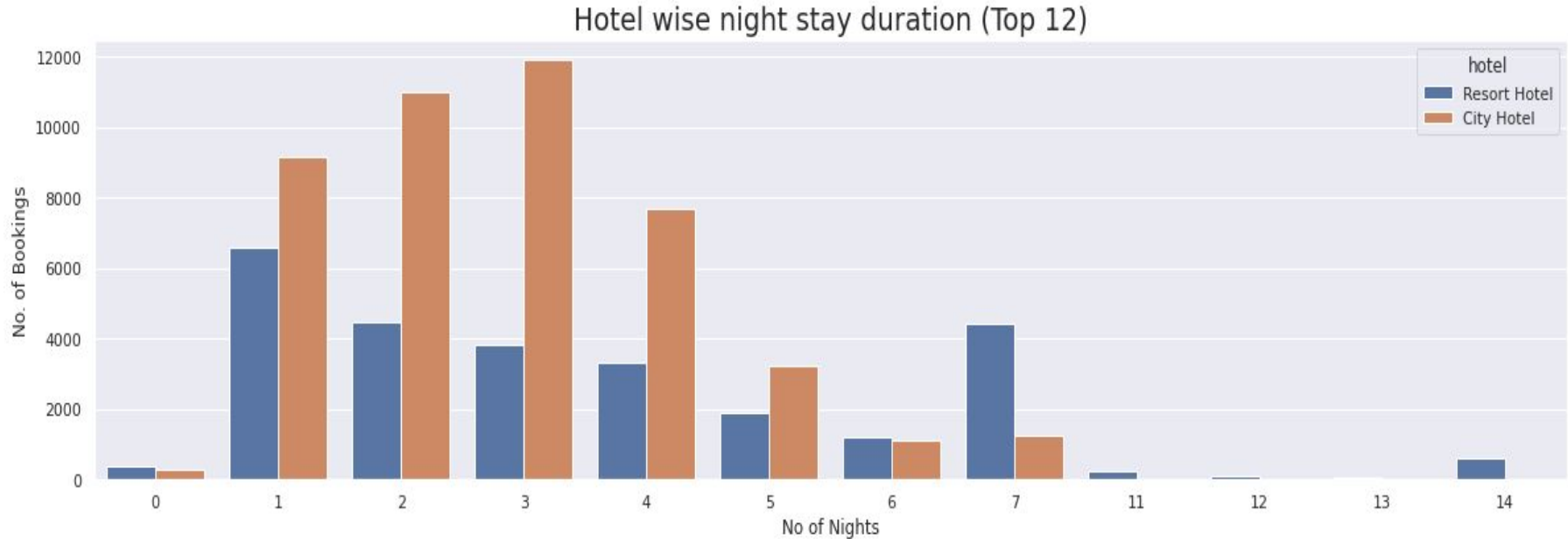
Bar chart shows there is more no bookings in city hotel in each year.

Distribution of Bookings Month-wise:



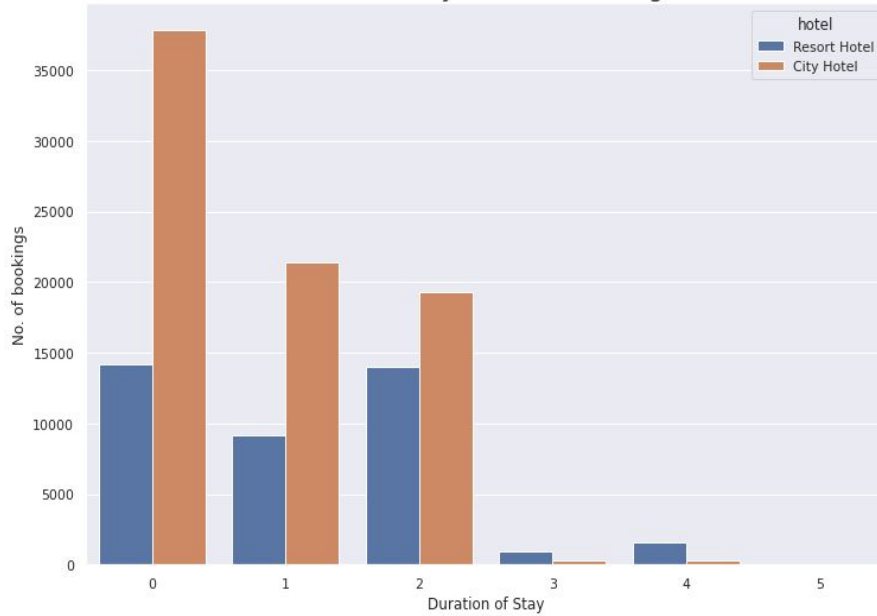
Monthly Distribution of booking in city hotel is also superior to Resort hotel.

Duration of Stay in Hotels:

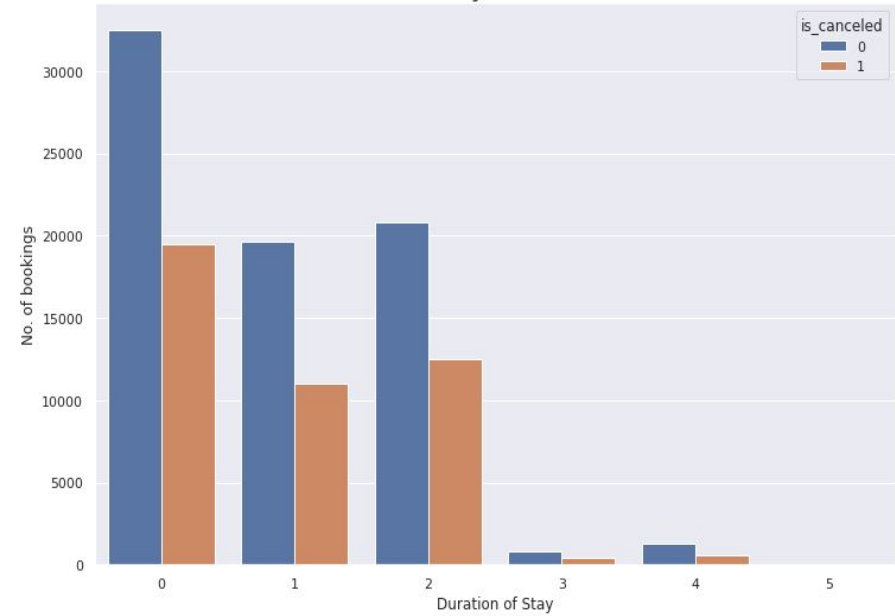


- For longer duration people booked resort hotel, where as for shorter duration people preferred city hotels over resort hotels.

Number of stays on weekend nights



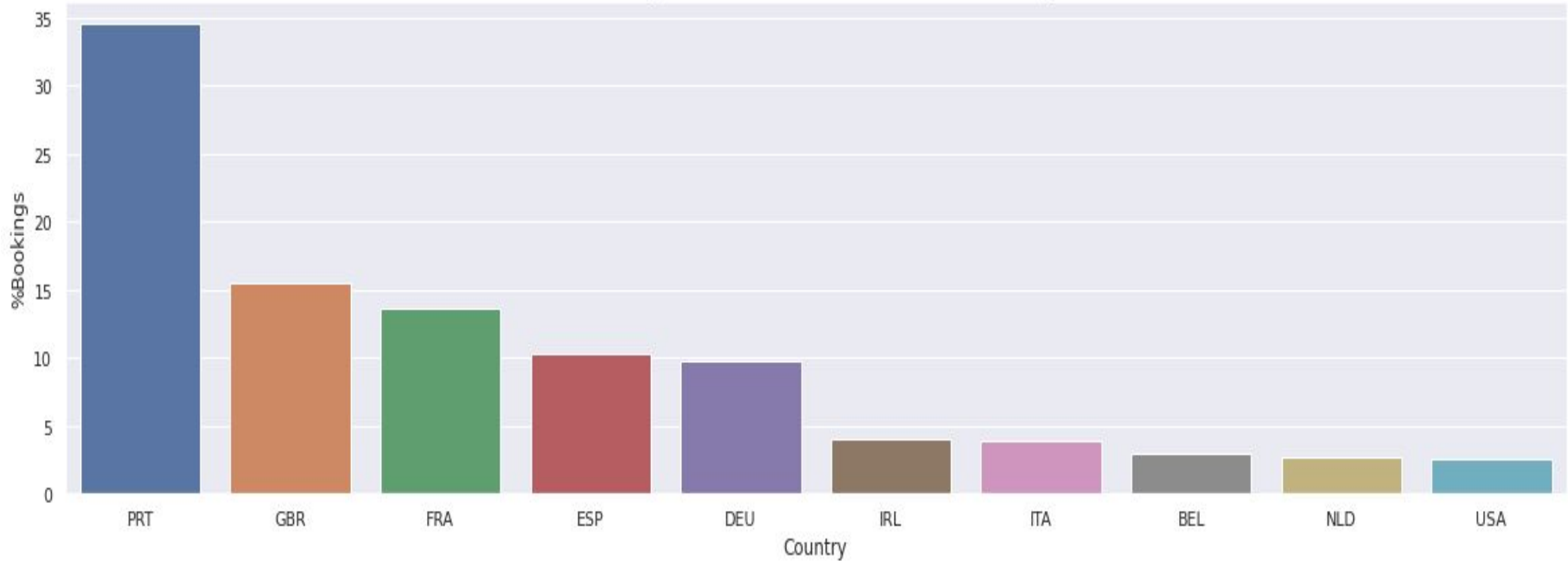
WeekendStay vs Cancellation



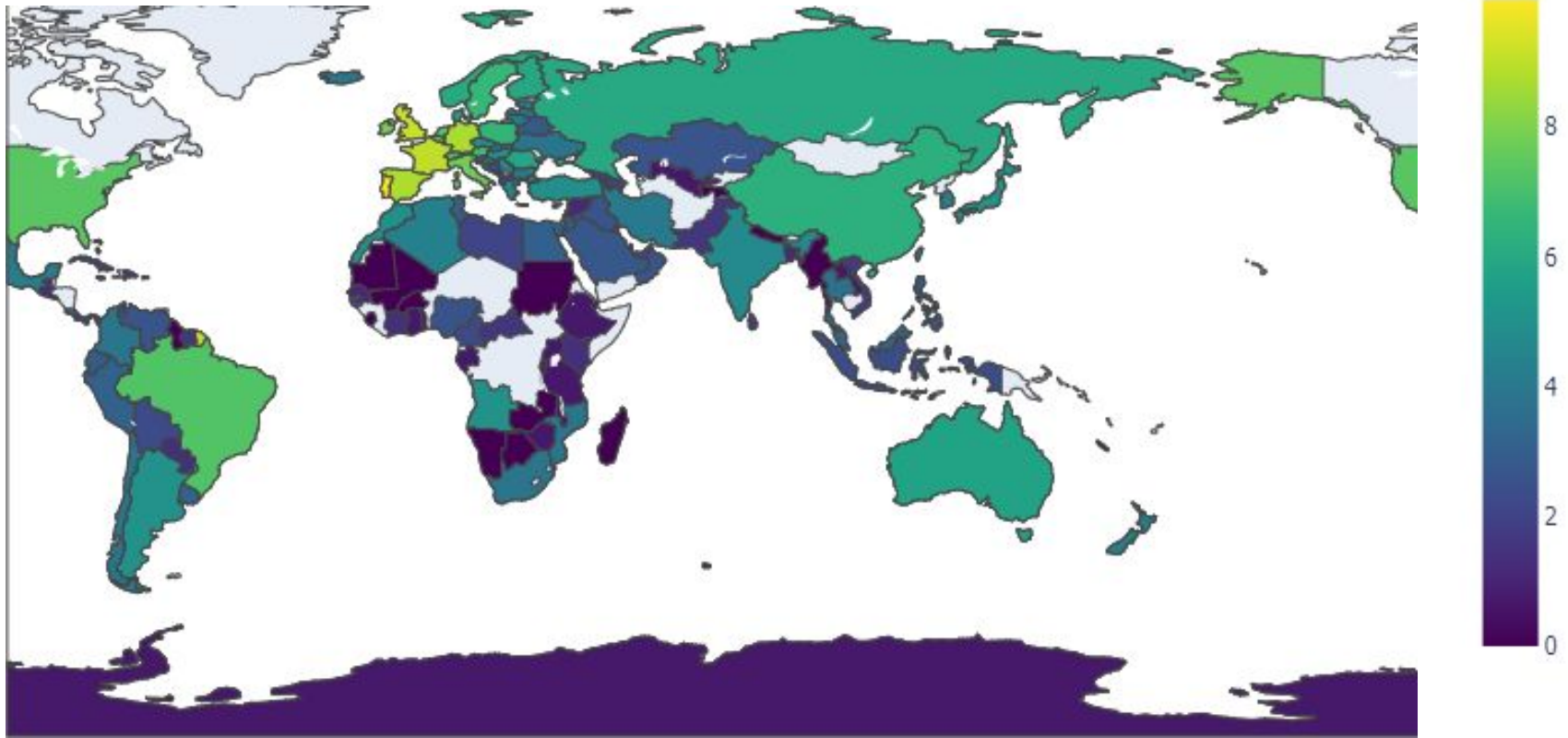
Weekend stays in both type of hotels Vs cancellations shows that cancellation of booking is directly proportional to overall booking numbers.

Geographical Analysis:

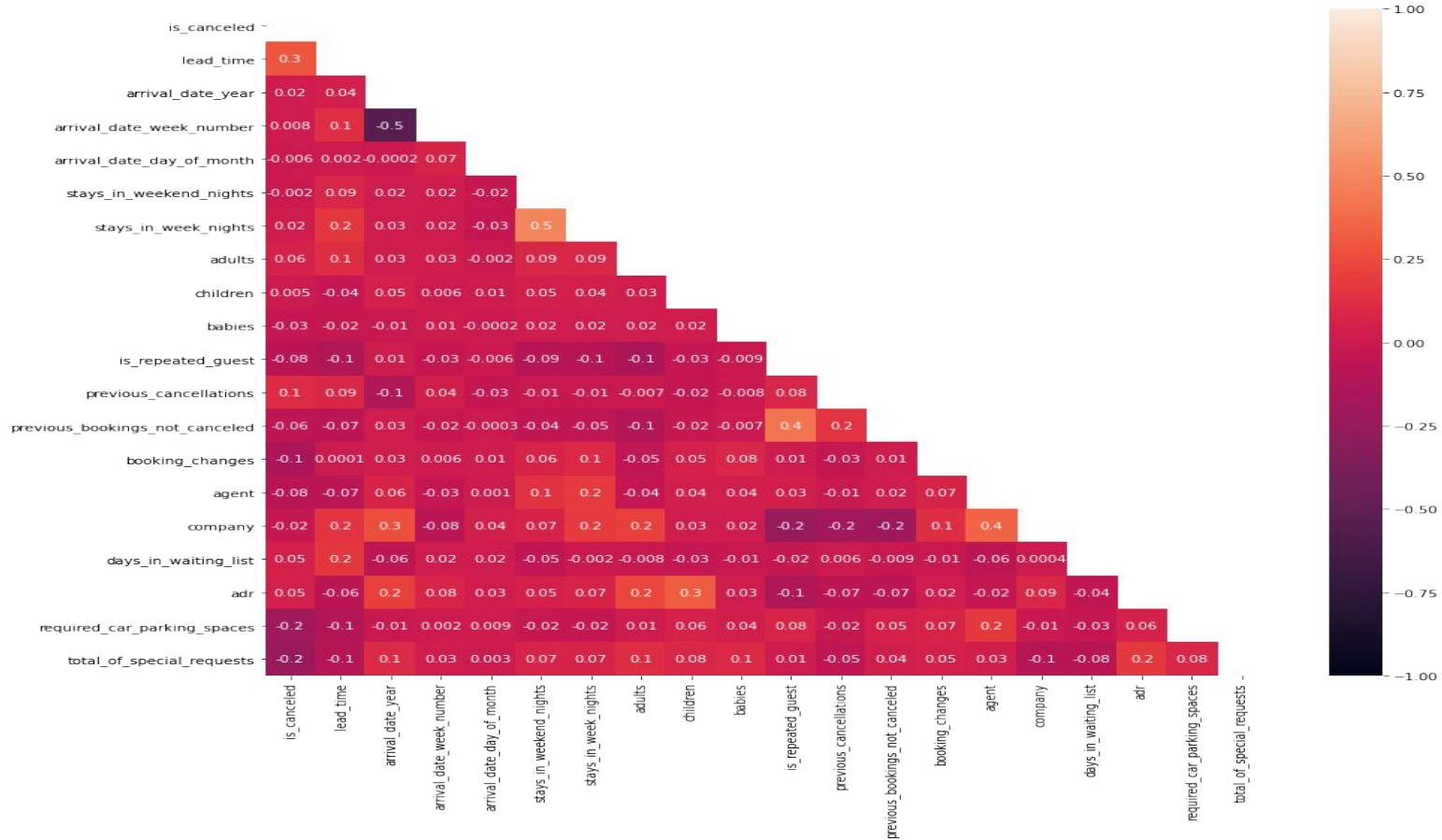
Country Wise Distribution of Bookings



In top-10 countries with highest no of visitors are from **Portugal**, lowest no of visitors are from USA

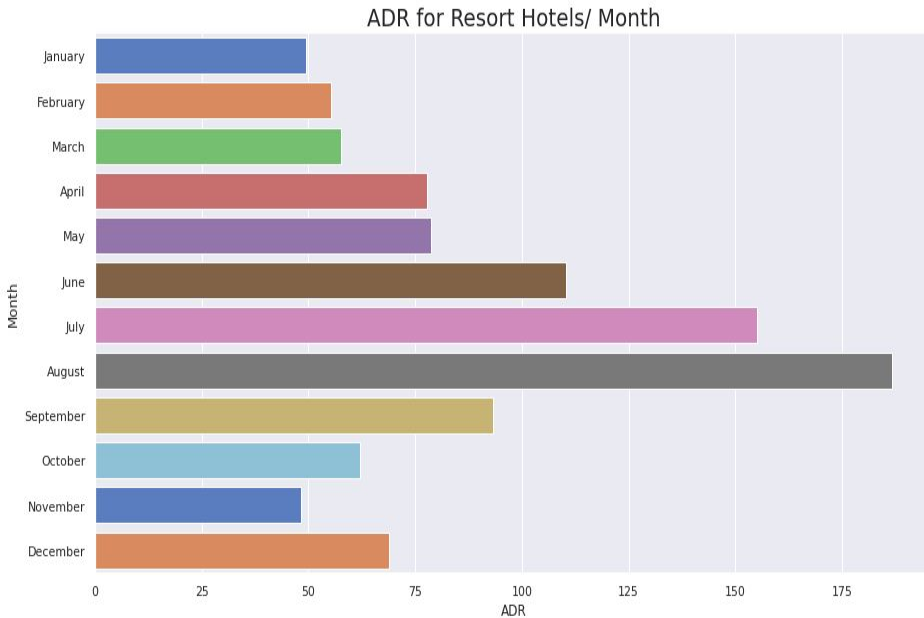
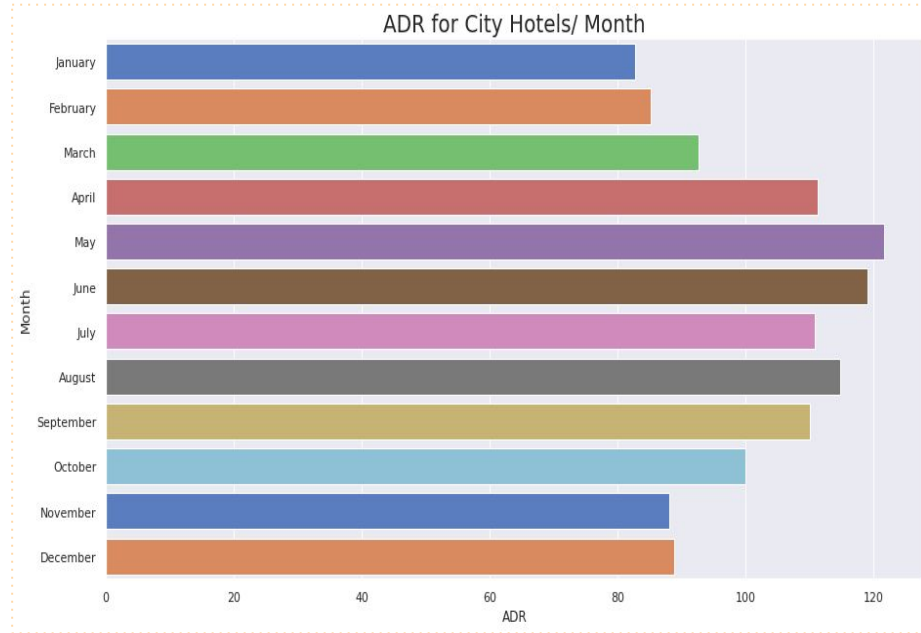


Darker(purple) part in the world map indicates less visitors and lighter(yellow) part shows more visitors



- If people are coming as a family and children are included then ADR of Hotel also increases and If any guest have stayed earlier also, then chance of booking getting cancelled decreases

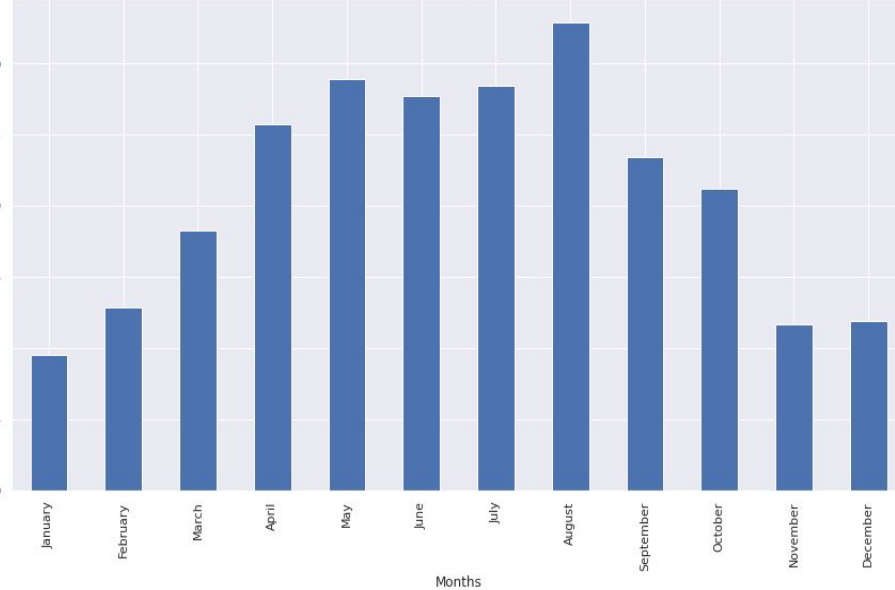
ADR Analysis



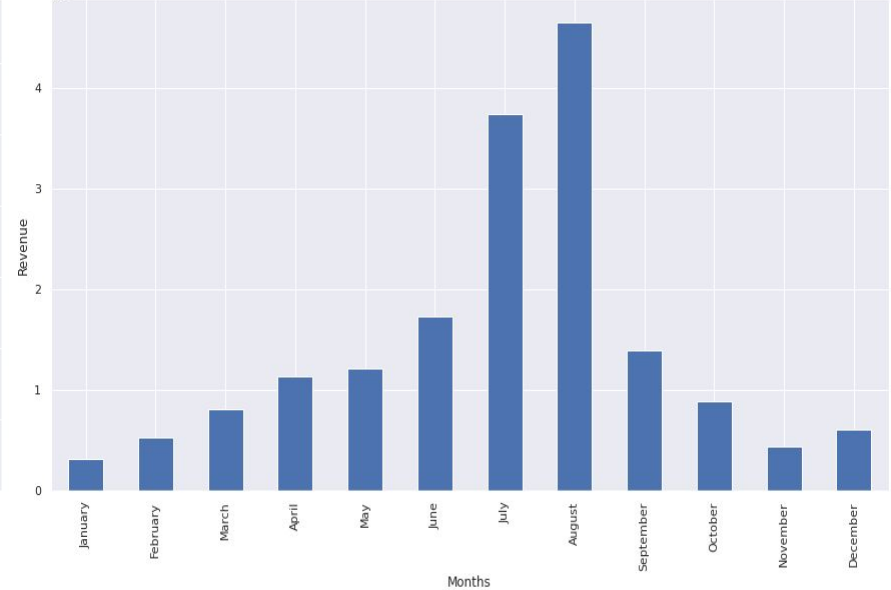
- Comparing ADR for both the hotel reveals, Resort hotel is having more ADR than city hotel, in the months: June, July and August.

Revenue Analysis

Revenue for City Hotel/ Month

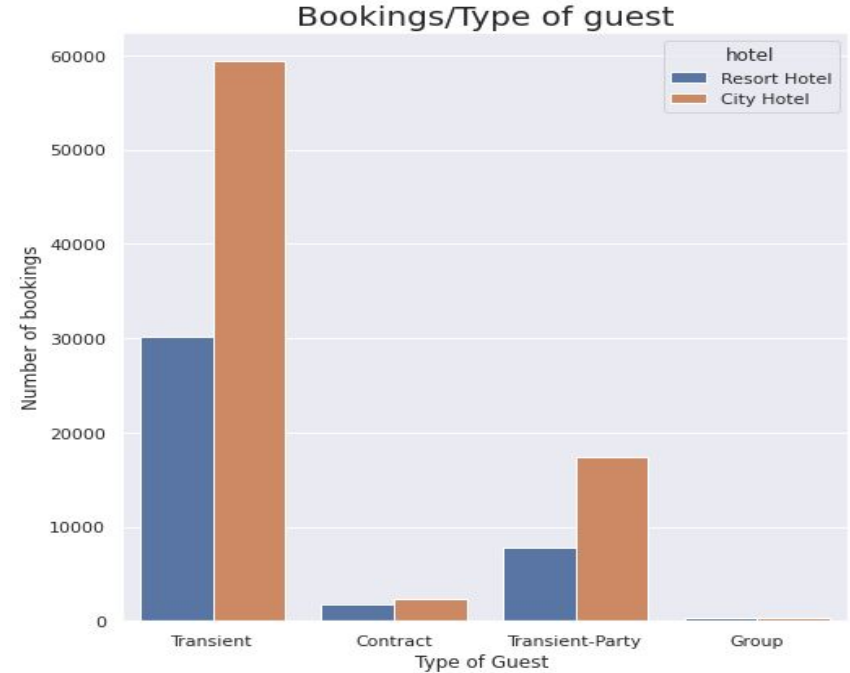
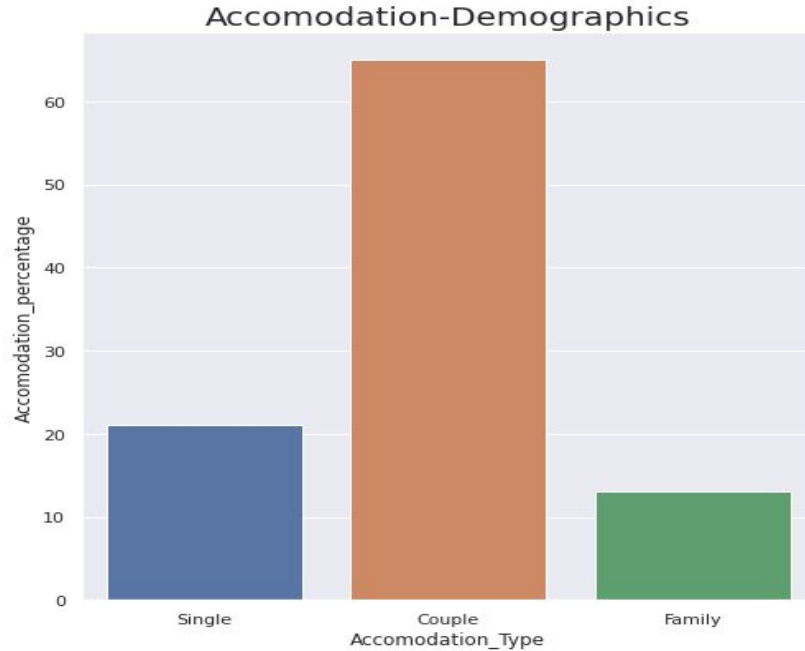


Revenue for Resort Hotel/ Month



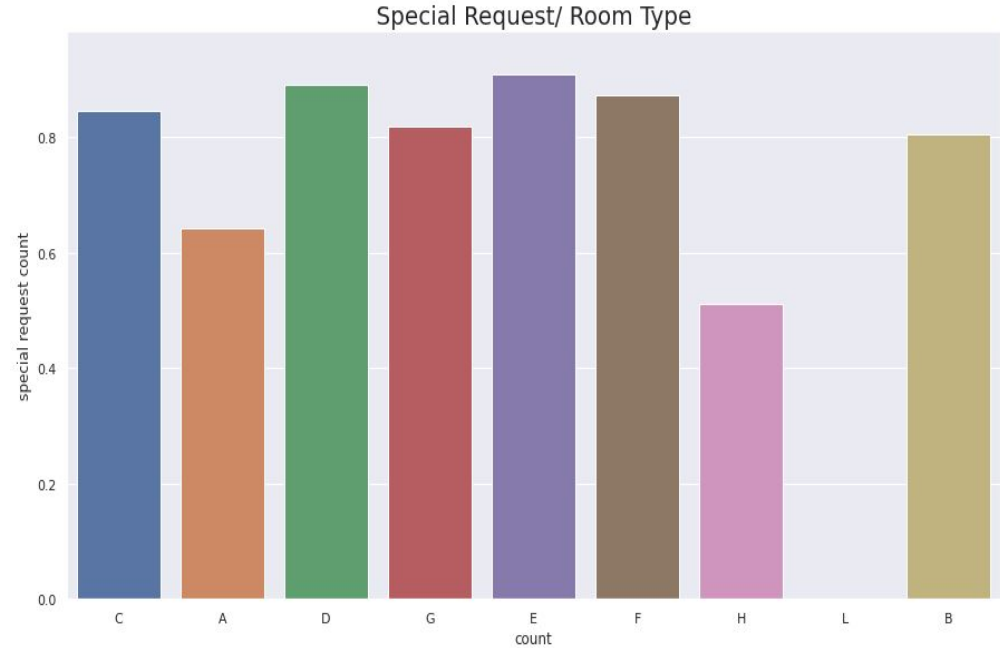
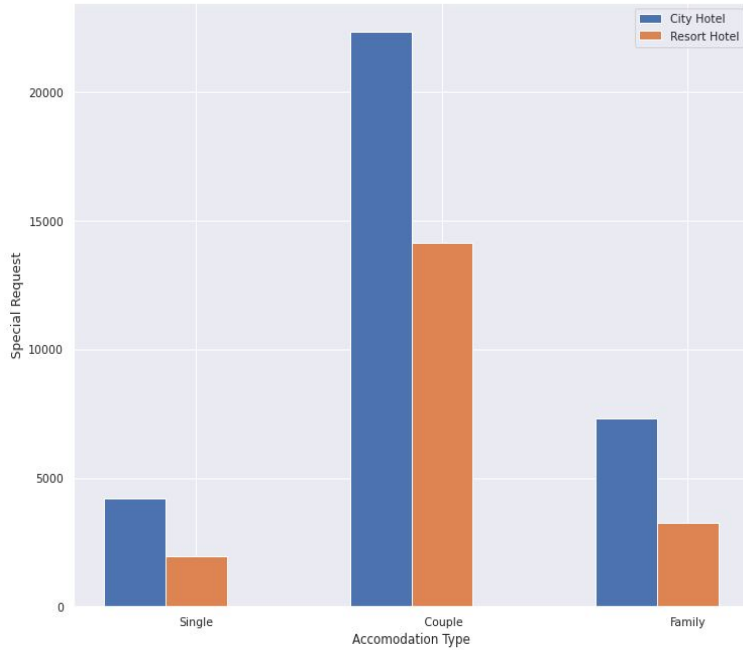
- Maximum revenue generated in the month of August and November- February(4 months) are not profitable for hotels as they observe least bookings during this period
- In Resort Hotels, July and August generates highest revenue.

Customer Analysis



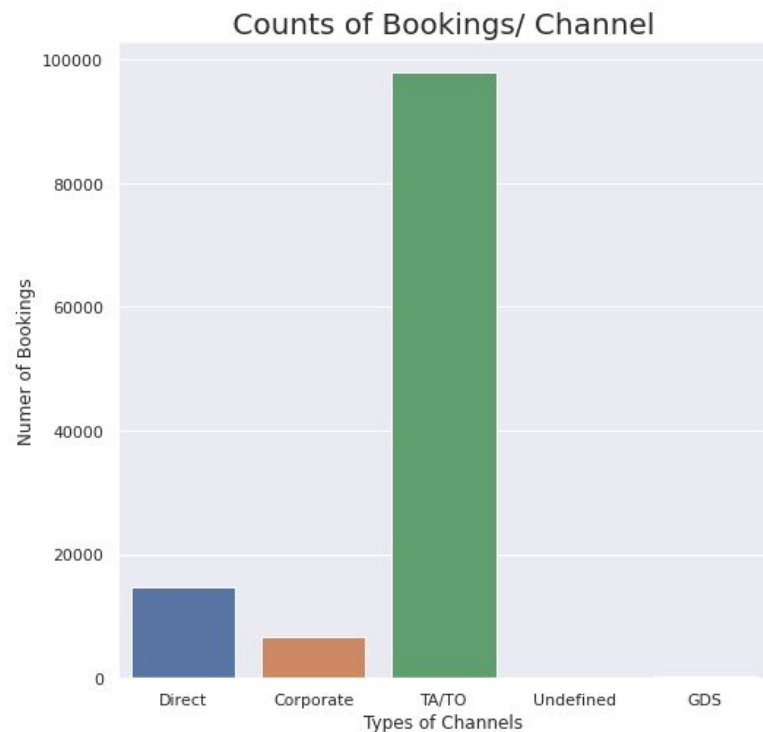
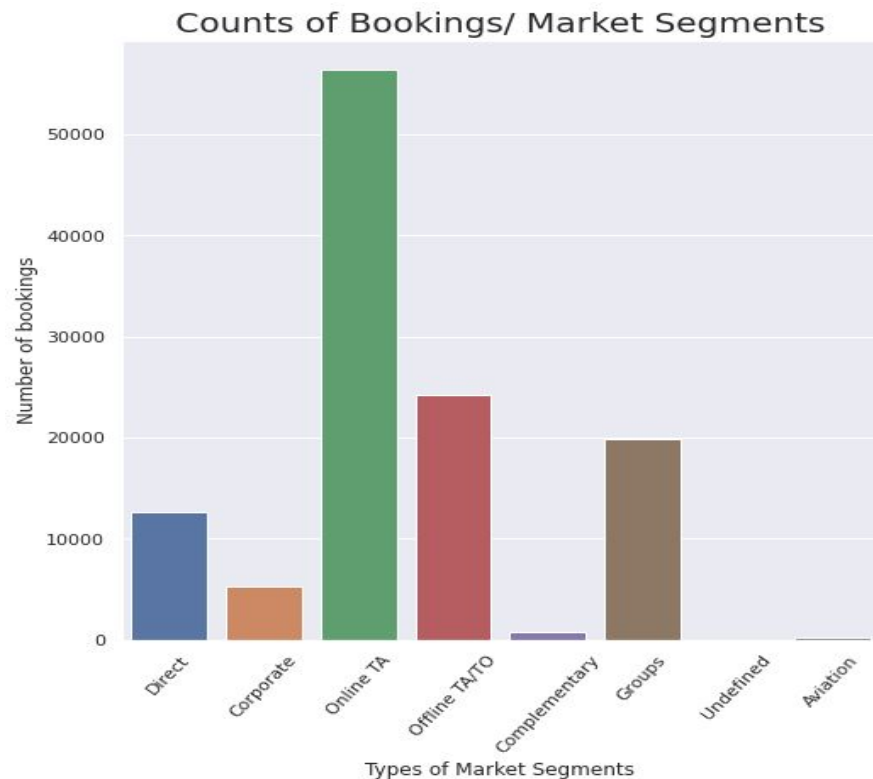
- Couple accommodation is 65%, singles accommodated 21-22%, and family 10-15% of totals.
- City hotel has highest no of booking w.r.t Transient/non-transient bookings

Special Requests



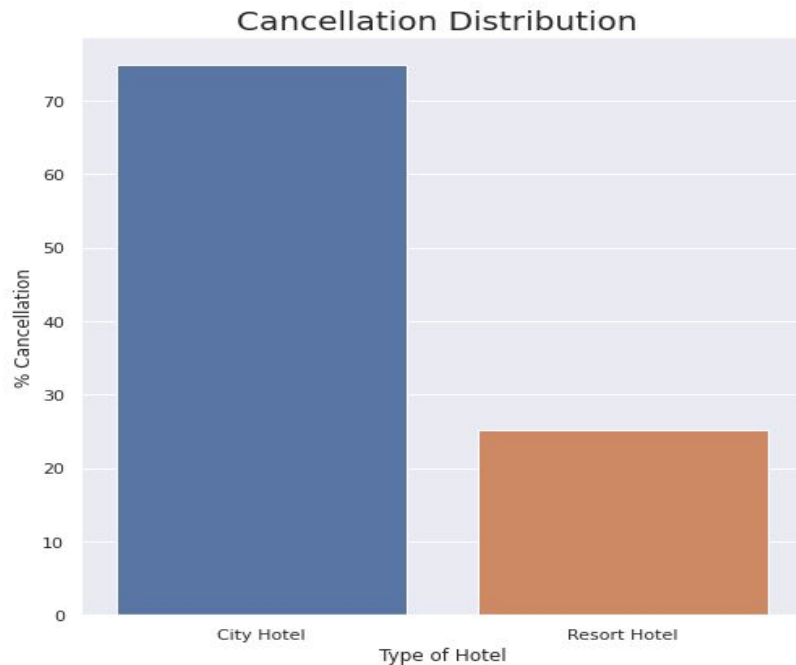
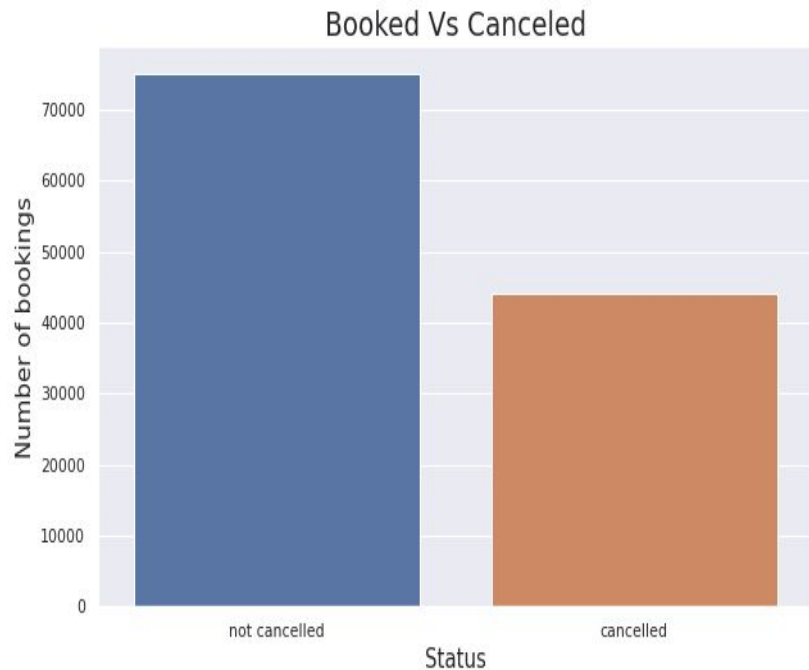
- Couples have made the most requests in both the hotels and singles have made the least requests.
- Room type A and H have somewhat less requests and room type L had the least .

Market Segment and Distribution channel

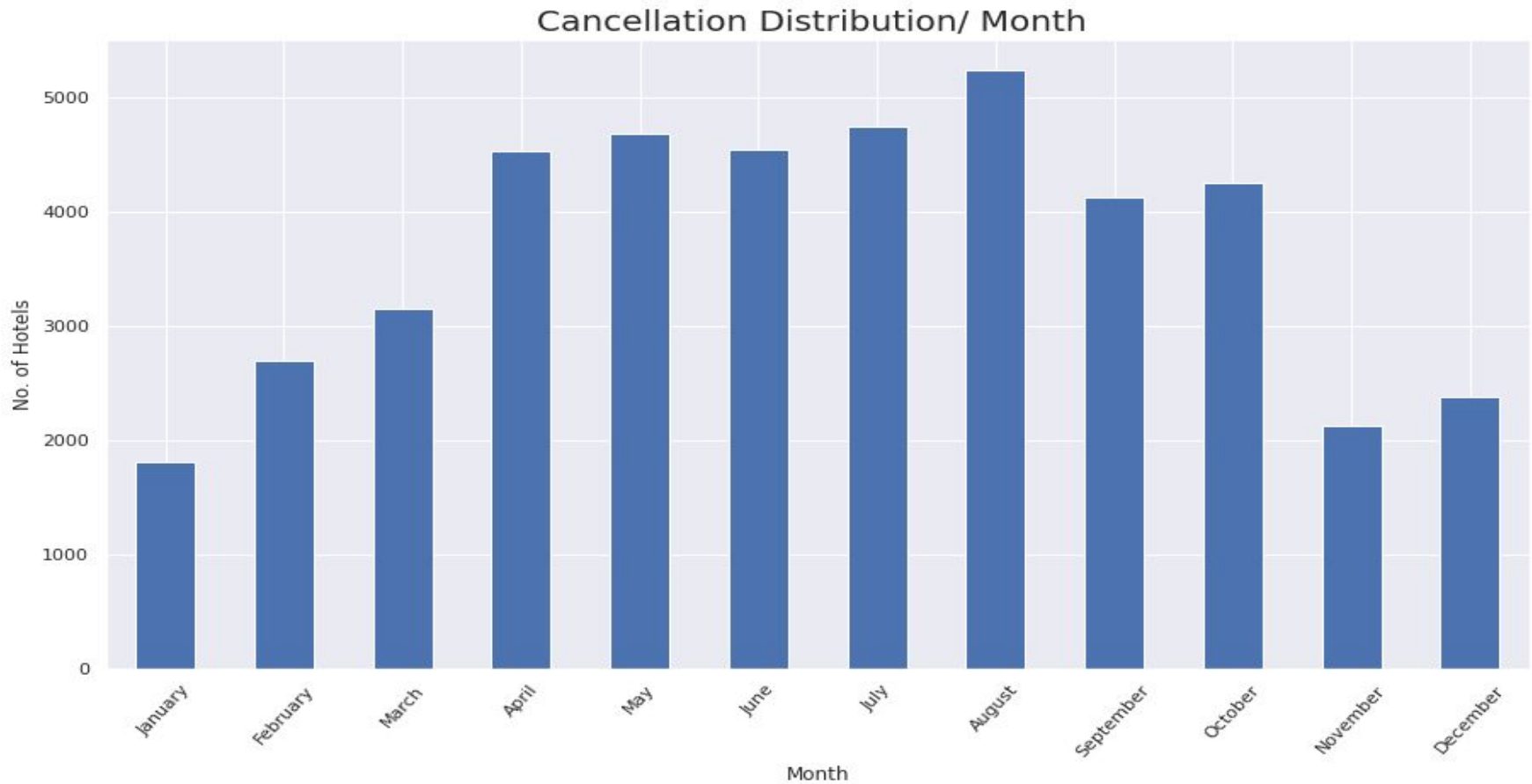


- TA/TO has highest no of booking count followed by Direct and corporate booking though distribution channel.
- TA brings most no of bookings in online TA followed by offline TA/TO, groups, direct and corporate.

Cancellation Analysis



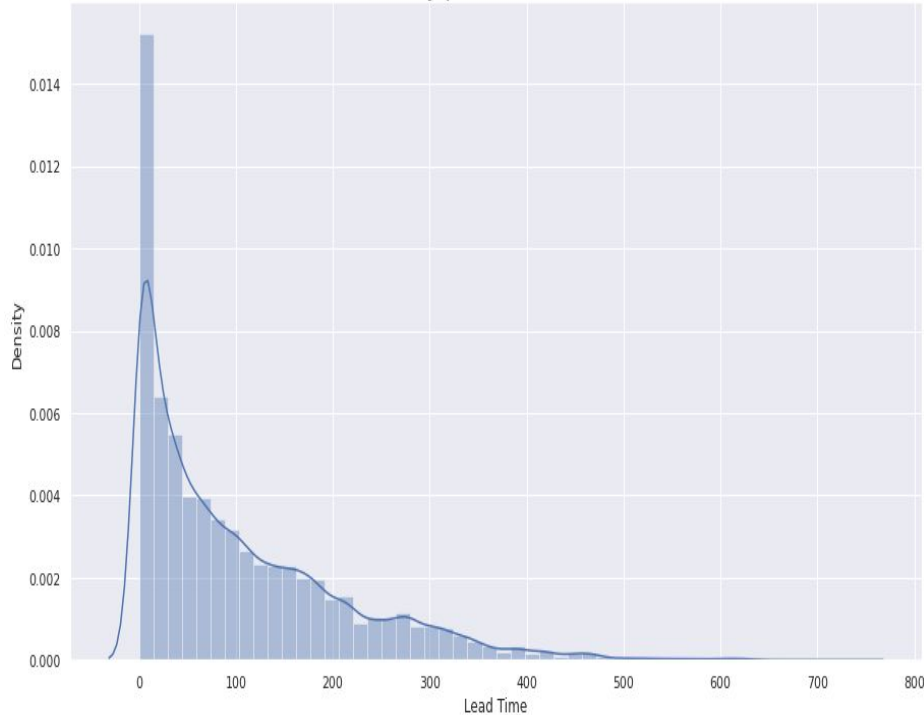
Most bookings & cancellations are from the City Hotel compared to Resort Hotel.



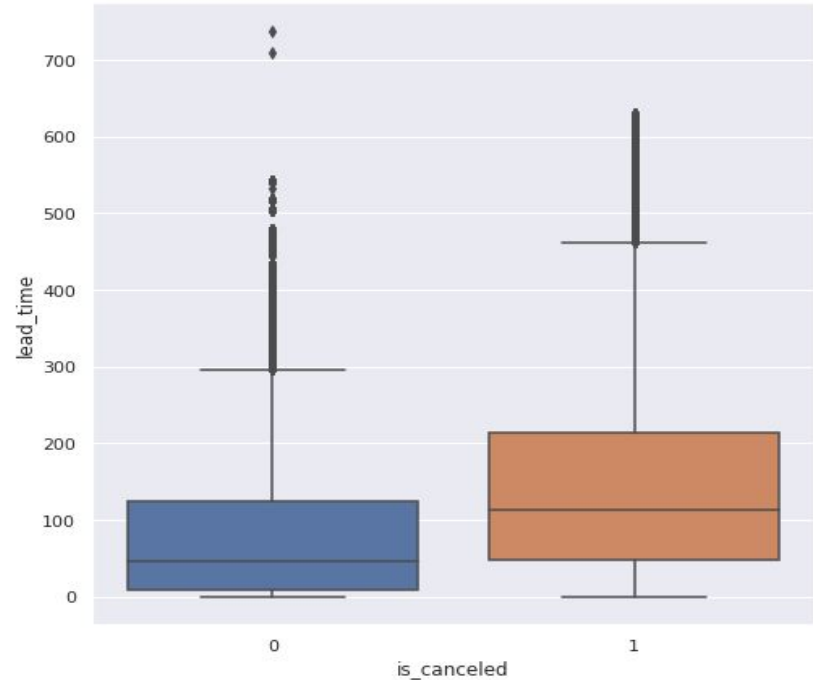
Cancellations are positively correlated with bookings, as we saw that bookings were highest in 2nd and 3rd quarter so is the cancellations.

Lead time Vs Cancellation:

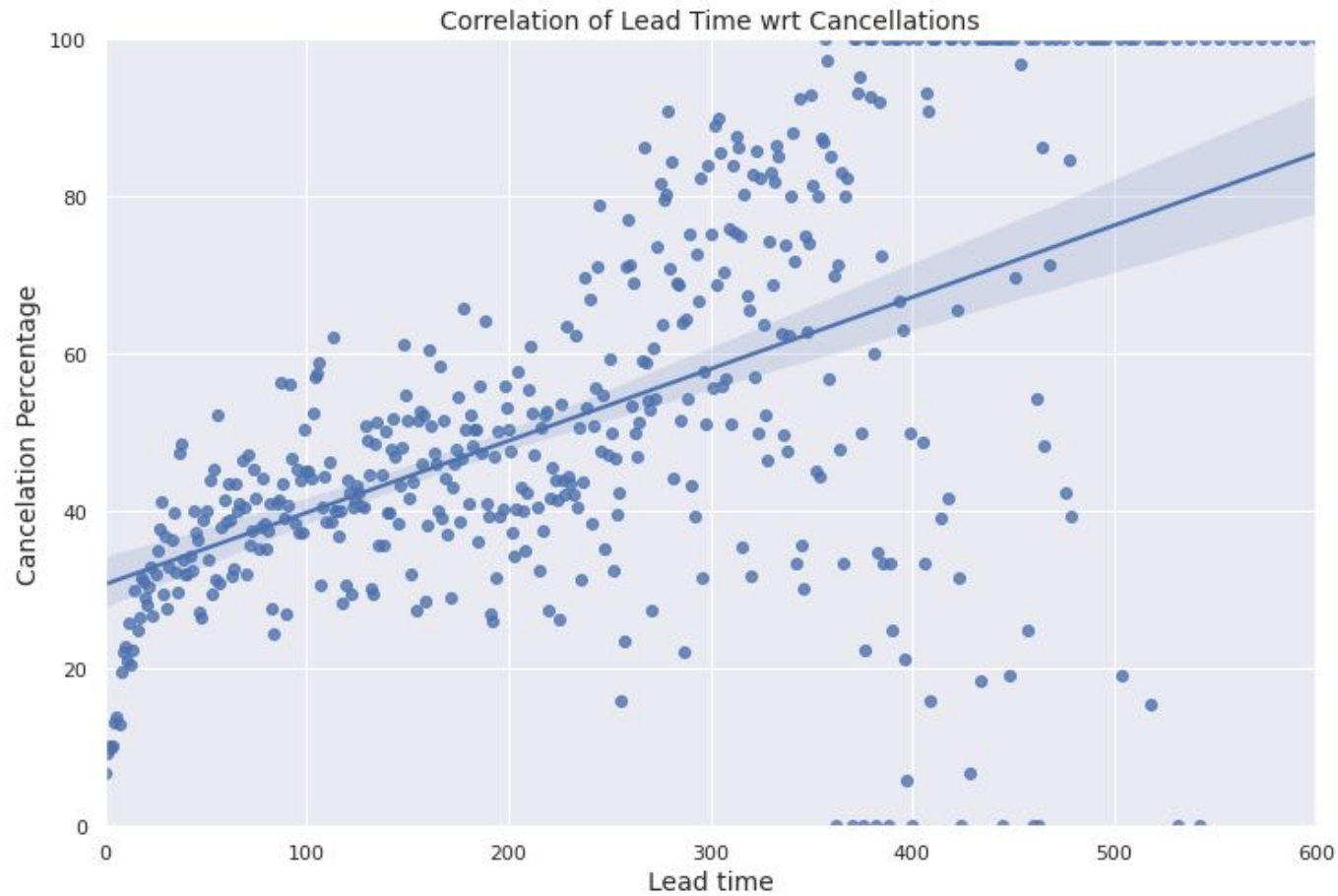
Density plot for Lead Time



Cancellation and Lead time

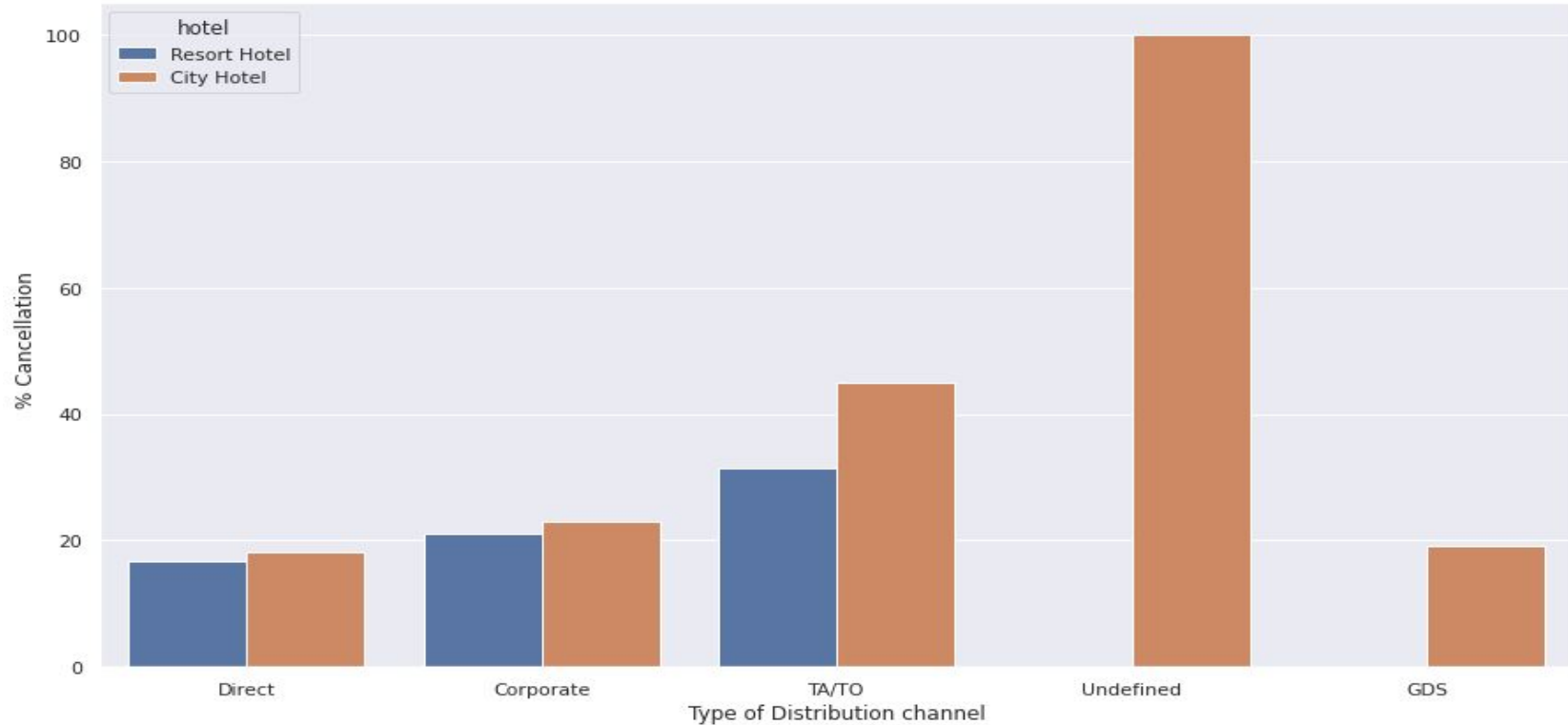


- Density plot shows that maximum bookings were done between 0-50 days prior to arrival.
- People who have done bookings 4 months prior to arrival have higher chances of cancellation compared to people who booked within a month

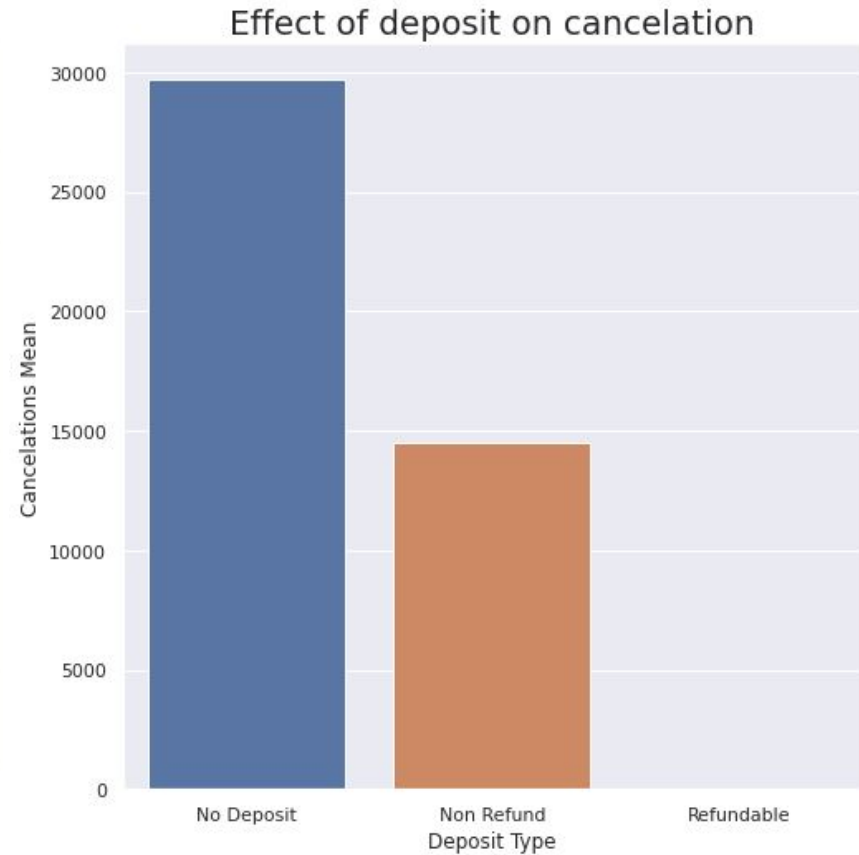
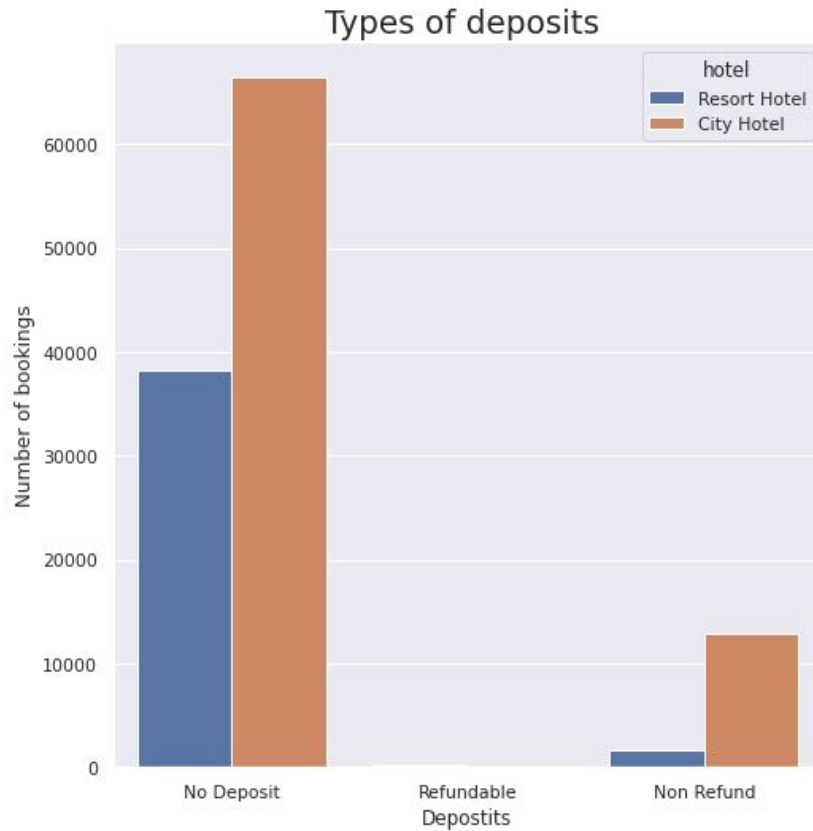


Correlation of lead time with respect to Cancellations - shows that with increase in lead time the no of cancellations also increase

Cancellation w.r.t. market segment



- The 'undefined' market segment favors the most no of cancellations for City Hotel while 'Groups' market segment favors the most no. of cancellations for Resort Hotel.



- Maximum bookings were done without any deposits, as customer paid only after arriving at hotel in these bookings.
- Maximum cancellation too was done in this type of bookings where customer did not paid anything in prior.

Conclusion:

That's it! We reached the end of our exercise. Time for Conclusion.

- How Many Booking Were Cancelled?
- What is the booking ratio between Resort Hotel and City Hotel?
- What is the percentage of booking for each year?
- Which is the busiest month for hotels?
- From which country most guests come?
- How Long People Stay in the hotel?
- Which was the most booked accommodation type (Single, Couple, Family)?
- When family's book their stay is the ADR affected ?

We observed that

- Almost 35% of bookings were canceled.
- More than 60% of the population booked the City hotel, which means majority of travellers prefer City Hotel over Resort Hotel.
- We have the highest booking % in 2016 i.e. around 47%. Also the second and third quarter results in most no. of bookings
- The busiest month for Hotels is August. The reason being that in European Countries August is a warm month and best for travelling.
- Portugal is the country from where the most no. of guests arrive and some of the other countries like Great Britain, France and Spain
- Most people stay for one, two, or three days.
 - a. Customer are staying in Resorts for longer durations, and in City Hotels for shorter durations.
 - b. We can arrange the packages of shorter duration for City Hotels to attract the guests
 - c. We can attract most no. of guest with some special service and special requests.
- Couple (or 2 adults) is the most popular accommodation type. Hotels can provide the discounts to couples as to engage them.
- Yes, ADR increases if family comes to stay. To attract Family accommodation, the hotel can provide the additional benefits and perks along with discount coupons